HUMAN BiOLOGY AND PUBLiC HEALTH JOURNAL OF THE AUXOLOGICAL SOCIETY

ISSN 2748–9957 | Vol. 2 2024 | pp. 1–10

# Tutorial: Statistics for Human Biologists – Introduction

Detlef Groth[1]

[1]University of Potsdam, Institute of Biochemistry and Biology, 14476 Potsdam-Golm, Germany

**Citation:**

Groth, D. (2024). Tutorial: Statistics for Human Biologists, Human Biology and Public Health 2. https://doi.org/10.52905/hbph2024.2.92.

Received: 2024-10-29
Accepted: 2024-11-19
Published: 2024-12-20

**Conflict of Interest:**

There are not conflicts of interest.

**Correspondence to:**

Detlef Groth
email: dgroth@uni-potsdam.de

**Keywords**

experimental design, dummy data, linear models, decision tree, programming language

## Abstract

For many researchers in the field of human biology, statistics is a foreign territory where they feel uncomfortable because of statistical vocabulary used and the many different ways of analyzing their data. In a series of five short review articles, I will try to help these researchers to understand the basic principles of data preparation, descriptive and inferential statistics to guide them in analyzing their data. The review series should be seen as a complement to the summer school lectures of the International Summer School of the University of Potsdam in Gülpe in the state of Brandenburg, Germany. In this first tutorial we discuss the role of statistics in research and the role why a statistical tool like R should be used and what are the main considerations before we actually should start analyzing the data.

**Take home message for students** Statistics requires careful experiment design and data preparation and is mainly determined by the data types of the variables of interest.

## Introduction

Statistics with its mathematical based terminology is a field that is often not well understood by researchers with a mainly biological background. At the Gülpe International Summer School at the University of Potsdam (Scheffler et al. 2024), which focuses on research in human biology such as the growth and development of children or the impact of different types of health problems on their maturation, we cover basic statistical techniques for analyzing single variables, relationships between two variables, and multivariate approaches. The final lecture also covers machine learning approaches such as linear models and decision trees.

In this first of a series of five tutorials, I will loosely follow these lectures, and I will promote the use of the statistical language R (R Core Team 2024) compared to more graphical oriented approaches like spreadsheets, explain the statistical workflow, and the main data types and structures which we need to represent our data within the analysis.

In subsequent sections we will cover basic types of sampling approaches to ensure representative selections of data, the different types of statistical analysis and then the basics of uni-, bi-, and multivariate statistics and at the end, machine learning.

## The scientific workflow

Normally, when investigating of a research question, the scientist should start with a testable hypothesis and then design experiments to investigate the question of interest (Fig. 1). After the experiments have been designed and carried out, data collection begins and the data are saved in various types of documents or databases. If the data collection takes a long time, or for illustrative purposes, the researcher can also create dummy data to prepare the analysis even before the data collection is complete. Such dummy data can even be used to evaluate the basic ability of the chosen analysis to find the desired effect if it were present within the experimental data. Thus, user generated dummy data can be used as a positive control if the expected effect is introduced into these dummy data or as negative control if the dummy data is created completely at random.

Experimental data usually needs to be processed before they can be used within our statistical software. This process, often referred to as data mining, is often tedious and prone to error, so that the capabilities to programming approaches are very helpful. It is not uncommon that most of the work on the software side is done in this step of data preparation and data import. During this process it is very important to check the data for obvious errors, for example sometimes height data is entered in cm, but there might be some values given in meters within the same column. So always check minimum and maximum of your data to identify possible errors in your data. This process of data checking is called data preprocessing, which is done before the actual analysis begins.

After the data are successfully delivered to the statistical software the actual statistical analysis begins, first describing the sample using numerical summaries and visualization techniques such as data plotting. Then comes the necessary process of generalization, after describing the sample we want to infer the population. We generalize from our sample data to the population. To do this, we need the mathematical background of probability theory to estimate the likelihood of observing effects like those we saw in the sample data compared to data free of the effect we are

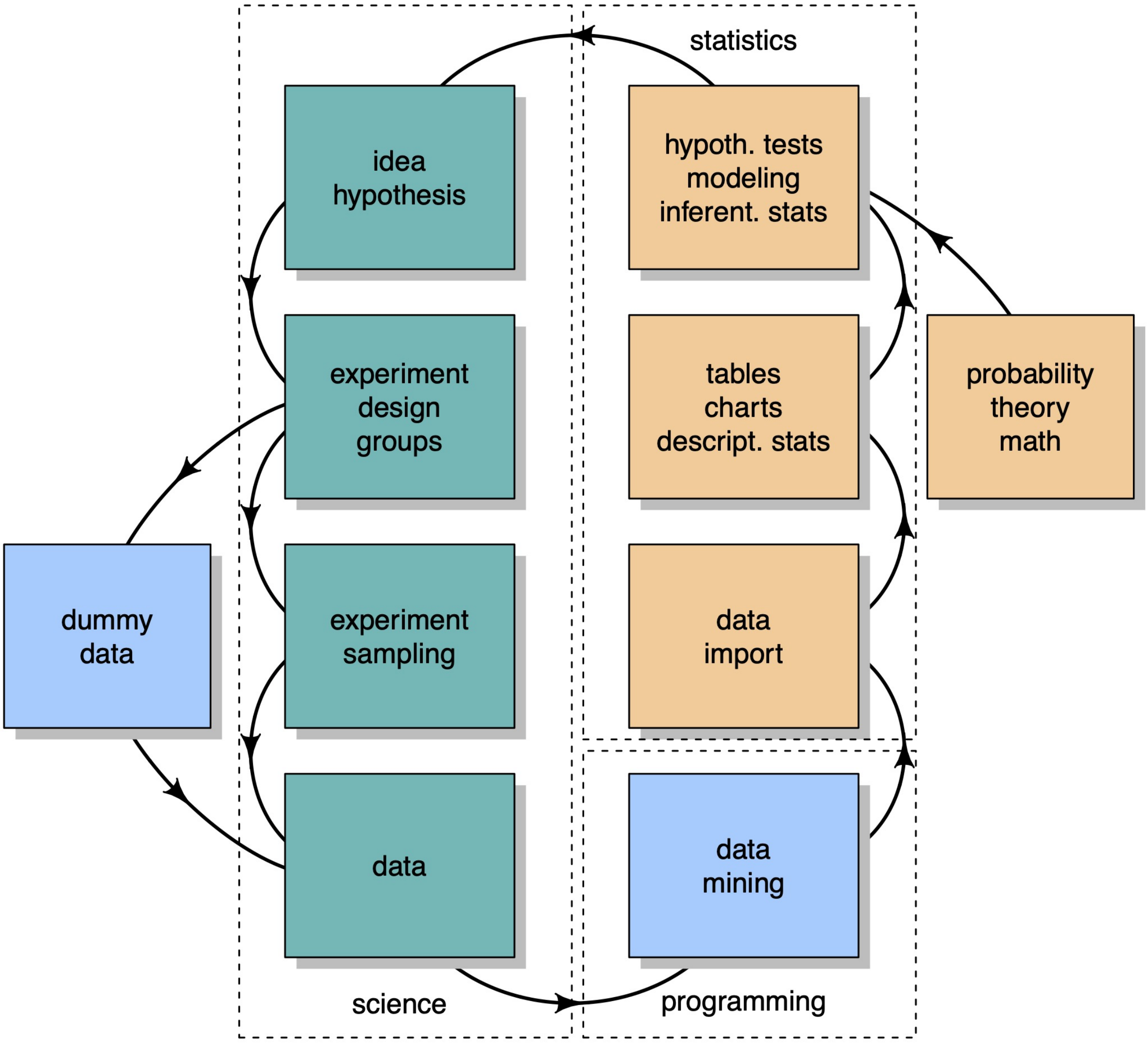


**Figure 1** The scientific workflow

investigating. We should then summarize our results and report effect sizes. If the system is well understood we can also start building models to explain our effect with some variables that allow us to make the appropriate inferences. Once the statistical part is done, new hypotheses often come to the researchers' mind and the process starts again.

## Comparing approaches to analyze data

There are two different approaches to performing statistical analysis, the first approach uses mainly graphical tools that provide menus, buttons, checkboxes, and similar interface elements to perform the required steps for analysis. The user interacts directly with the data and solves possible problems by visually editing the data. This type of analysis is mostly performed using spreadsheet programs like Microsoft Excel or LibreOffice Calc.

The main disadvantage of this more graphical approach is that there is no separation of data and statistics. Both are done within the same file, with calculations and visualizations embedded next to the data. Re-running the analysis with new or updated data at a later stage is often prone to error. Most often, old analysis files are then used as templates and the new data is copied over the old data. However, if the dimensions of the new data are different, fewer or more rows or columns, parts of the analysis may be lost or simply wrong. Greater problems occur when the data is spread across more than one file, sometimes even dozens or hundreds. Repeating the same analysis on all these files is often impossible.

In contrast programmatic approaches using statistical software tools such as the aforementioned R, Python (Python Software Foundation, https://www.python.org/), or Matlab (The MathWorks, Inc., Natick, Massachusetts, United States) separate data and analysis. Typically, the data is kept separate from the analysis. Script files operate on these data and should also deal with reformatting or renaming issues as well as fixing problems in the data within the script files and not within the data itself. So, the researcher is guided to not edit the data anymore, but to make the necessary data adjustments within the analysis program. This approach provides easy reuse of the analysis for other data of the same type or if the data is updated.

## Using R in statistics

Besides Python and Matlab the statistical programming language R is the most widely used tool to analyze data using the programmatic approach. Since R is free and open-source software and available for all major platforms, we use it for our summer school as the main tool to analyze the student's data. All the different languages have their advantages and disadvantages, but we prefer R here because of its open-source license and due to the fact that R's focus is statistics: help facilities about functions and packages are more unified and easier to access for the user.

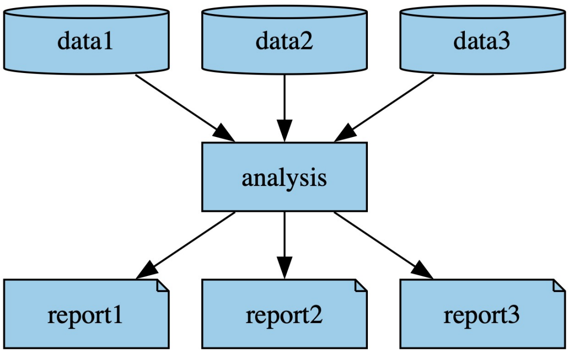
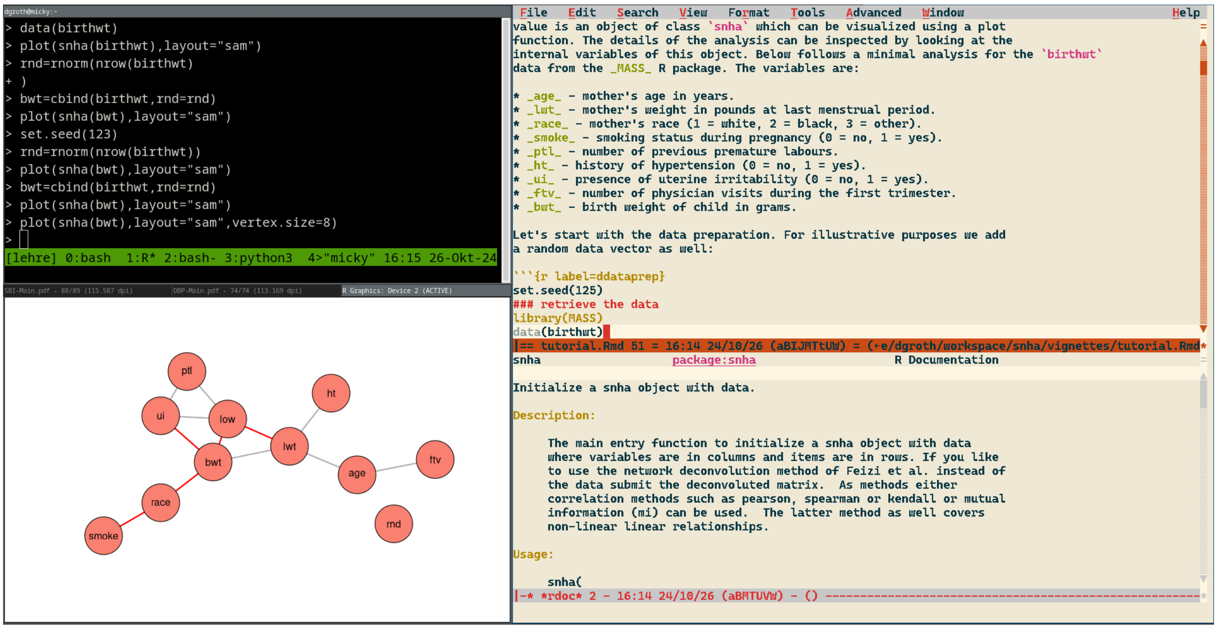


**Figure 2** Programmatic approach – the same analysis script using an R, Python or Matlab script, can be applied to many data files (Excel, CSV, Database files) to perform the same analysis for all the different input data.

For R, a variety of tools exists to simplify the analysis process for the user. A typical R programmer's workspace consists of a text editor window where the analysis is written and stored in a script file, an interactive R console and a plot window where visualizations are placed (Fig. 3). Some more advanced users may use their own text editor and run the R console and other windows or terminals needed to organize the analysis. Less experienced users often use software solutions which provide similar workspace features within one application, but these solutions are often more resource-intensive, such as the widely used RStudio (RStudio Team 2024).


dgroth@micky:~
SBI-Main.pdf - 88/89 (115.587 dpi)
DBP-Main.pdf - 74/74 (113.169 dpi)
R Graphics: Device 2 (ACTIVE)
ptl
ht
ui
low
lwt
bwt
age
ftv
race
rnd
smoke
File
Edit
Search
View
Format
Tools
Advanced
Window
Help

```
> data(birthwt)
> plot(snha(birthwt),layout="sam")
> rnd=rnorm(nrow(birthwt)
+ )
> bwt=cbind(birthwt,rnd=rnd)
> plot(snha(bwt),layout="sam")
> set.seed(123)
> rnd=rnorm(nrow(birthwt))
> plot(snha(bwt),layout="sam")
> bwt=cbind(birthwt,rnd=rnd)
> plot(snha(bwt),layout="sam")
> plot(snha(bwt),layout="sam",vertex.size=8)
> 
[lehre] 0:bash  1:R* 2:bash- 3:python3  4>"micky" 16:15 26-Okt-24
```

~~~
value is an object of class `snha` which can be visualized using a plot
function. The details of the analysis can be inspected by looking at the
internal variables of this object. Below follows a minimal analysis for the `birthwt`
data from the _MASS_ R package. The variables are:

* _age_ - mother's age in years.
* _lwt_ - mother's weight in pounds at last menstrual period.
* _race_ - mother's race (1 = white, 2 = black, 3 = other).
* _smoke_ - smoking status during pregnancy (0 = no, 1 = yes).
* _ptl_ - number of previous premature labours.
* _ht_ - history of hypertension (0 = no, 1 = yes).
* _ui_ - presence of uterine irritability (0 = no, 1 = yes).
* _ftv_ - number of physician visits during the first trimester.
* _bwt_ - birth weight of child in grams.

Let's start with the data preparation. For illustrative purposes we add
a random data vector as well:

```{r label=ddataprep}
set.seed(125)
### retrieve the data
library(MASS)
data(birthwt)
|== tutorial.Rmd 51 = 16:14 24/10/26 (aBIJMTtUW) = (←e/dgroth/workspace/snha/vignettes/tutorial.Rmd*
snha                     package:snha                    R Documentation

Initialize a snha object with data.

Description:

     The main entry function to initialize a snha object with data
     where variables are in columns and items are in rows. If you like
     to use the network deconvolution method of Feizi et al. instead of
     the data submit the deconvoluted matrix.  As methods either
     correlation methods such as pearson, spearman or kendall or mutual
     information (mi) can be used.  The latter method as well covers
     non-linear linear relationships.

Usage:

     snha(
|-* *rdoc* 2 - 16:14 24/10/26 (aBMTUVW) - () ---------------------------------------------------*
~~~

**Figure 3** Typical workspace of an R programmer performing an analysis. In the upper left, there is an interactive R console where the user can interactively perform the analysis steps. On the lower left there is a plotting window for the visualizations performed within the R console. On the right is the user's text editor, in this case it is a Jasspa MicroEmacs session (https://github.com/bjasspa/jasspa), but any other editor based on the user's preferences could be used instead. Here an analysis is performed for the snha package – the St. Nicolas House analysis (Groth D. et al. 2019; Hermanussen et al. 2021).

## Scientific question and hypothesis formulation

As shown in Figure 1, research should be driven by concrete scientific questions and testable hypotheses. Project descriptions such as: "We are investigating the influence of nutrition and education on growth and health" would be better formulated as specific questions such as: "Does the parental education have an influence effect on children's height?" followed by a directed, testable hypothesis. Here an example: "We hypothesize that the education level of the parents is positively associated with children's height". Giving a direction such as "positively associated" should be preferable to an undirected hypothesis such as "We hypothesize that there is a significant association between parental education and children's height". In summary, we need a clear scientific question and a testable hypothesis that can either be verified or rejected.

The entire process of creating a testable hypothesis could be outlined in a very brief form as follows:

1. determine your relevant variables
2. formulate a general undirected hypothesis and a null hypothesis H0
3. specify the direction of the effect to formulate a directed hypothesis which serves as your alternative hypothesis H1
4. draw a figure of your hypothesis to illustrate it
5. ensure that your hypothesis can be tested

The null hypothesis (H0) is usually the hypothesis that we want to reject, it is for example the hypothesis that there is no association between two variables, like parents' education and children's height are independent of each other. In contrast, the alternative hypothesis (H1) is the hypothesis we want to accept, for example that parents' education and children's body height are (positively) related.

## Research types

There are two basic types of research, correlational research where we observe what happens in our sample without the goal or ability to manipulate any of the variables of interest. Case control studies are of this type of research, you compare two groups and go back in time to compare these groups with regards to specific risk factors. An example for such a study conducted in the 1980s where data from a New York veterinary hospital that showed that the incidence of injury and mortality of cats peaked for falls of about seven to eight stories and then declined for falls from greater heights (Diamond 1988). Because the veterinarians did not perform experiments, for ethical reasons, they could only study cats that were being cared for at the hospital. Such studies can be subject to bias, for example, in this case not all cats that might have died immediately could be included because they were not admitted to the hospital. In the terminology of a case control study, groups can for instance be seen as dead or alive, the risk factor was story height. An example of a student's summer school project is the study of Kotnik et al. (Kotnik et al. 2024) which determined secular trends in anthropometric characteristics such as body height, fat mass, fat-free mass, and external skeletal robustness.

The other type of research is experimental research, in which the researcher can manipulate a variable, such as a pharmacological study in which one group receives a drug and the other receives placebo.

This type of research is generally preferred because the use of appropriate sampling strategies can ensure a reliable comparison of the two groups. In contrast to correlational studies, purely observational studies can be seriously flawed. For example, it is known that there is healthy vaccinee bias in older age groups (Hoeg et al. 2023; Fireman et al. 2009), which may to lead to an overestimation of vaccine effectiveness. In contrast, younger individuals are more likely to have an opposing vaccination bias, because younger individuals with illnesses are more likely to be vaccinated than healthy individuals. In summary, it may be appropriate to choose experimental research when possible.

## Sampling

When conducting your research, make sure that your sampling procedure is truly random. For example, if you do a survey, ensure that selecting one item does not change the probability of selecting another item. For example, you should not ask the person you are interviewing to bring another person to the same interview. This would lead to a biased sample.

For example, to ensure truly random sampling, the following techniques could be used:

- Rolling a dice, then count the individuals or items and based on the dice points select the next sample. This process is called simple random sampling: each individual has an equal chance of being selected
- Take every nth sample, for example, every 10th, but be careful not to bias the sample: if you select households and every 10th house is a house at a street corner, this process is called systematic random sampling.
- Clustered sampling ensures that each category proportionally represents the population regarding sex, age, ethnicity etc. Within these clusters the other two techniques mentioned above can be used

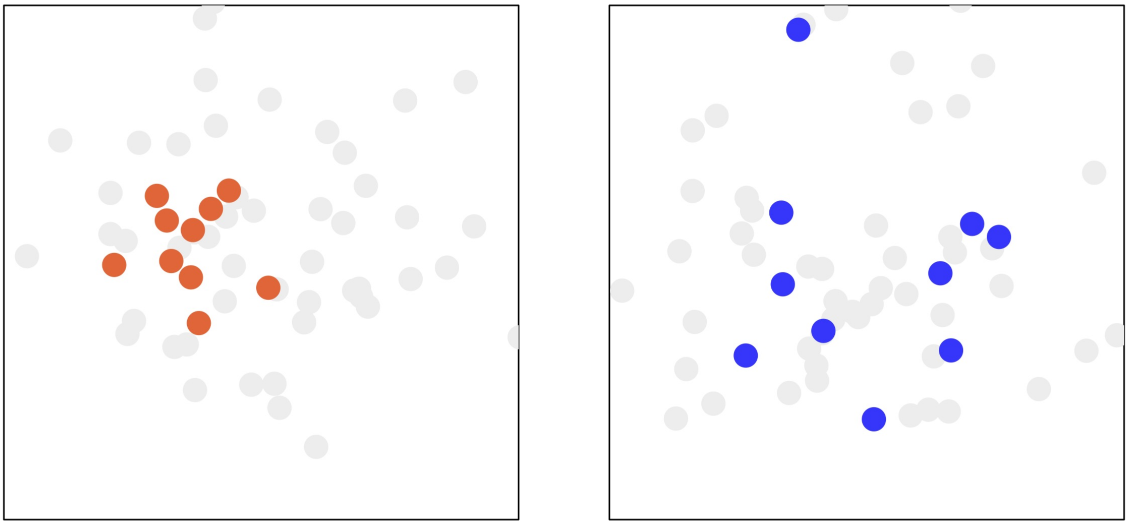


**Figure 4** Biased sample (orange dots) on the left from a population (gray dots) that may be the result of non-random sampling, unbiased sampling on the right (blue dots) that should better represent the population.

Without these techniques, you cannot expect your samples to be truly random, e.g., in the greenhouse, you might intuitively choose plants that grow taller because they are easier to reach.

If you want to compare two groups, start by randomly selecting the sample items, create pairs of items with similar characteristics, such as similar age, similar health status, same sex, etc., and then randomly place one of the two items in group 1 and the other in group 2. This ensures that both groups have very similar characteristics.

Biased sampling maybe unavoidable. Take a clinical trial: usually you start with hospitalized volunteers, who however might be less healthy than the overall population.

Another well studied example in this context is the observed higher mortality rate of patients admitted to hospitals on weekends, which is related to the generally lower admission rate on weekends, with only the most urgent cases being admitted compared to weekdays, leading to a higher mortality rate (Meacock et al. 2019).

## Data handling advice

The data you collect is usually stored in files or databases. Most users will save and use files, often files that can be read by spreadsheet programs such as Microsoft Excel or LibreOffice Calc. To efficiently handle these files as input for your analysis, some standard rules should be followed.

First, make a backup of your data from the first version you receive. Backup your data regularly, for example adding the date to the file name in case you change or fix the data. Give your data reliable, short column names that do not contain spaces or special symbols. If you want a legend to explain the column names, create a separate sheet in the same file for the legend, with the short name in the first column and the long explanation in the second column.

If you are using a programming language like R or Python, you should not modify the data file, but keep it in the unmodified form that you get from your collaborator or your machine. Fix column names, data problems etc. within your R or Python script, not within the data file. If you get an updated data file you can simply change the input file name. So, the rules for statistical programming languages are as follows:

- Leave the data file unchanged.
- Fix column names and data problems within your R script file.
- After making these corrections in an R script file, you start your analysis in the same or in another R script file.
- Updates are easy. If the file format does not change, just switch the data filename in your script file.

## Data types

Once you have collected your data, the main points that will guide your statistical analysis methods are the number of variables you are examining and the type of variables you are using in your analysis. The data type of a variable can be divided into quantitative and qualitative types. Numerical types (quantitative data)- such as age, body height or z-score of age - can follow a particular distribution and range of values. Qualitative data such as sex (male, female) or smoking status (yes, no) might be differentiated into categorical types which still have a specific order, for example low, medium, or high, or categorical types which have no order (sex, ethnicity) and boolean types which can be seen often as yes/no values (or 1 or yes, 0 for no). For descriptive and inferential statistics there are furthermore different methods of data summaries, visualizations and testing in dependence with the appropriate data types.

## Data structures

When you analyze your data in an R script, you need to put that data into some data structures. Table 1 lists the most important data structures and their properties for the programming languages R and Python. The data structures can be divided either into structures that can only store the same type of data, such as only numbers; or structures that can store different types of data simultaneously, such as numbers and text in the same variable.

These are typical examples for data structures: Vectors contain sequences of values that are all of the same type. Matrices are two-dimensional, but can store only values of a single type. Arrays are similar to matrices, but can have more than two dimensions. These data structures are usually accessed by positional indices.

In R and Python, structures that can store different types of data are called lists. Each element of a list can again contain different types of data, even other lists. These structures are very flexible for storing data, although they are computationally less efficient because the memory addresses for the different values do not follow regular intervals based on a given data type. The data frame can be thought of a special kind of list in a two-dimensional form where each column is a vector and all vectors have the same length.

### Box 1: Terminology

- **alternative hypothesis (H1)** – usually hypothesize that there are changes, differences between two groups, or that there is a dependency between two variables
- **cluster sampling** – sampling ensuring proper presentation of existing population groups, for instance based on sex, age, etc. in the sample
- **correlational research** – observation driven research of what happens without intervention of the researcher like in case control studies
- **experimental research** – type of research where variables are manipulated like in a clinical trial with drug and placebo groups
- **null hypothesis (H0)** – usually hypothesize that there are no changes, differences

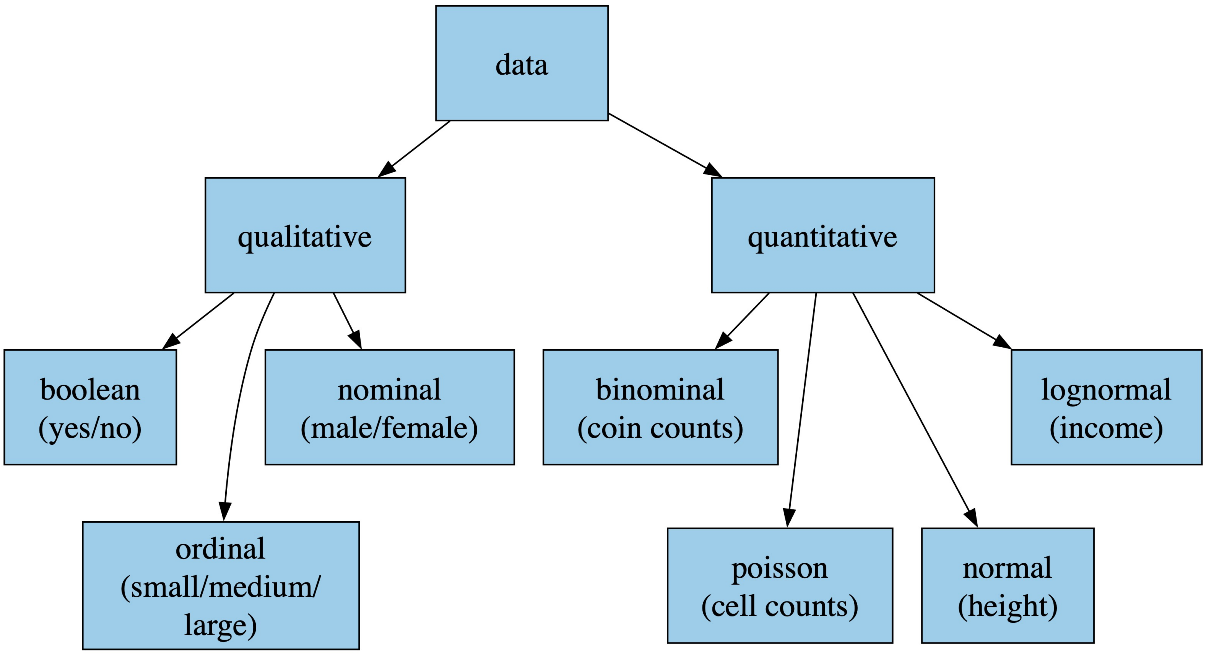


**Figure 5** Statistics is controlled by data types. We can basically distinguish between categorical/qualitative (left) and numeric/quantitative (right) data types. Here are some of the main data types, with examples in parentheses.

between the compared groups, or that two variables are independent

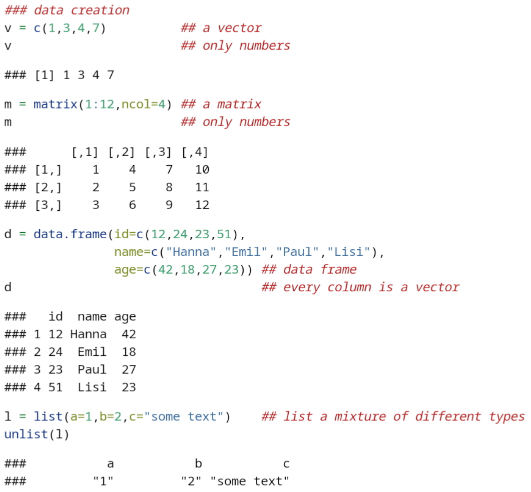
- **selection bias** – due to non-random sampling techniques some groups are overrepresented in the sample in comparison to the population
- **simple random sampling** – every element has the same chance of being selected, implemented for instance using a dice
- **systematic random sampling** – sampling with intervals, for instance surveying every 10th person

## Outlook

In this tutorial we covered basic considerations for performing representative sampling of data, why to use a programming language like R for statistics, and the basic data types and structures which are available for the programming languages R and Python. In the next tutorial of this series, we will discuss methods of descriptive and inferential statistics for univariate analysis and explain statistical terms such as P value and confidence intervals.

**Table 1** The basic data structures and functions for creating them for the R and Python programming languages.

| data structure | data types | dimensions | R function | Python function |
|---|---|---|---|---|
| vector | 1 | 1 | c | array.array |
| matrix | 1 | 2 | matrix | array.array |
| array | 1 | >= 2 | array | array.array |
| list | >= 1 | different | list | list / tuple |
| dictionary | >= 1 | different | list | dict |
| data frame | >= 1 | 2 | data.frame | pandas.DataFrame |



```
### data creation
v = c(1,3,4,7)          ## a vector
v                       ## only numbers

### [1] 1 3 4 7

m = matrix(1:12,ncol=4) ## a matrix
m                       ## only numbers

###      [,1] [,2] [,3] [,4]
### [1,]    1    4    7   10
### [2,]    2    5    8   11
### [3,]    3    6    9   12

d = data.frame(id=c(12,24,23,51),
               name=c("Hanna","Emil","Paul","Lisi"),
               age=c(42,18,27,23)) ## data frame
d                                  ## every column is a vector

###   id  name age
### 1 12 Hanna  42
### 2 24  Emil  18
### 3 23  Paul  27
### 4 51  Lisi  23

l = list(a=1,b=2,c="some text")    ## list a mixture of different types
unlist(l)

###           a           b           c
###         "1"         "2" "some text"
```

**Figure 6** R sample code for creating the data structures, vector (v), matrix (m), data frame (d) and list (l).

## Acknowledgements

I would like to thank the students and lecturers of the International Student Summer School in Gülpe at the University of Potsdam for their support and to the Summerschool KoUP Funding of the University of Potsdam, Germany. I would also like to thank Steven Phillips, one of the developers of MicroEmacs, who helped me in preparing MicroEmacs to provide the features that make it a powerful text editor for R, Python and Markdown.

## References

Diamond, J. M. (1988). Why cats have nine lives. Nature 332 (6165), 586–587.

Fireman, B./Lee, J./Lewis, N./Bembom, O./van der Laan, M./Baxter, R. (2009). Influenza vaccination and mortality: differentiating vaccine effects from bias. American Journal of Epidemiology 170 (5), 650–656.

Groth D./Scheffler C./Hermanussen M. (2019). Body height in stunted Indonesian children depends directly on parental education and not via a nutrition mediated pathway-Evidence from tracing association chains by St. Nicolas House Analysis. Anthropologischer Anzeiger, 445–451. https://doi.org/10.1127/anthranz/2019/1027.

Hermanussen, M./Assmann, C./Groth, D. (2021). Chain reversion for detecting associations in interacting Variables – St. Nicolas House Analysis. International Journal of Environmental Research and Public Health 18(4), 1741. https://doi.org/10.3390/ijerph18041741.

Hoeg, T. B./Duriseti, R./Prasad, V. (2023). Potential 'Healthy Vaccinee Bias' in a study of BNT162b2 vaccine against Covid-19. New England Journal of Medicine 389 (3), 284–286. https://doi.org/10.1056/NEJMc2306683.

Kotnik, K. Z./Golja, P./Pikel, T. R. (2024). Secular trends in anthropometric characteristics and their associations with external skeletal robustness among Slovenian young adults population. Human Biology and Public Health 1.

Meacock, R./Anselmi, L./Kristensen, S. R./Doran, T./Sutton, M. (2019). Do variations in hospital admission rates bias comparisons of standardized hospital mortality rates? A population-based cohort study. Social Science & Medicine 235, 112409.

R Core Team (2024). R: a language and environment for statistical computing. Vienna, Austria 2024. Available online at https://www.R-project.org.

RStudio Team (2024). RStudio: integrated development environment for R. Boston, MA 2024. Available online at http://www.rstudio.com/.

Scheffler, C./Hermanussen, M./Groth, D. (2024). 6th International Student Summer School on 'Human Growth: Data Analyses and Statistics'. Human Biology and Public Health 1.